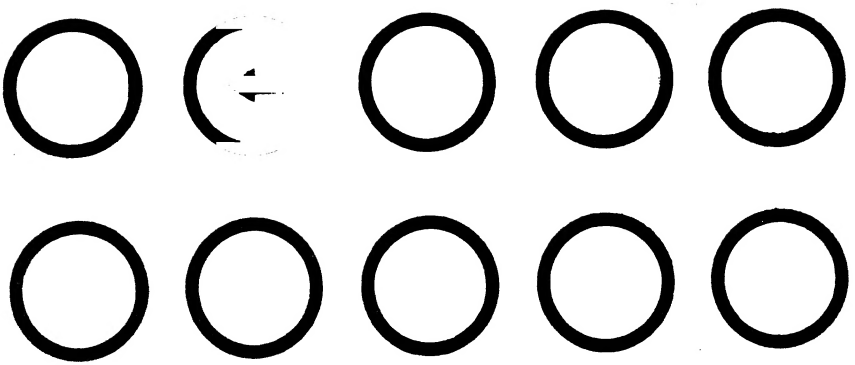
RESEARCH ABSTRACTS



RESEARCH CELL
ASSOCIATION OF INDIAN UNIVERSITIES
NEW DELHI

The paper used for printing this book was made available by the Government of India at a concessional rate.

Price : Rs. 6.00

Published by the Association of Indian Universities Rouse Avenue, New Delhi and Printed at Pearl Offset Press 5/33 Kirti Nagar, Industrial Area, New Delhi 110015.

PREFACE

The Examination Research Cell (ERC) of the Association of Indian Universities has been from time to time investigating into various fundamental aspects of university examinations,Internal Assessment, Grading, Test and item analysis, Practical Examinations to name only a few. Results of these research projects have been reported in the form of Monographs some of which have been revised to include experiences of teachers/ colleges/universities during implementation, in subsequent editions.

At the same time, certain research studies have been conducted and it is felt that these have to be reported in the form of Research Abstracts and three such abstracts are now getting ready.

In this second Research Abstract, the following in-depth studies have been included.

1. On the Reliability of Ratings in Interviews
2. Question Choice in Examinations
3. Simplified Methods of Test and Item Analysis
4. A comparative study of Reliability of Objective type examinations
5. A comparative study of Reliability of Choice type examinations

It is hoped that teachers, papersetters, examiners and other will find the research studies and the results and conclusions will be very helpful and useful and that they will be guided to better evaluation of their students' performance.

Constructive suggestions for advancing some of these studies, will be most welcome.

New Delhi.
22nd July, 1978.

V. Natarajan

ACKNOWLEDGEMENTS

The Examination Research Cell is indebted to the financial support from the Ministry of Education and Social Welfare, Government of India in particular to Mr. S. N. Pandita, Joint Secretary.

The Cell is grateful to Dr. L. S. Negi, President, AIU; Dr. Amrik Singh, Vice-Chancellor, Punjabi University; Shri I. J. Patel, Vice-Chancellor, Gujarat Agricultural University; Dr. A. K. Dhan Vice-Chancellor, North Eastern Hill University for their guidance and advice.

Mr. Ved Prakash, Research Assistant; Miss Sudha, Statistician; Mr. Ramakrishnan & Miss Veena, Assistants, Mr. V. K. Chugh, Steno and Mr. Vinod, IBM Typist have all helped in bringing out these abstracts.

Mr. Vijay Batra of Pearl Offset Press has helped in printing these abstracts.

It is not possible to acknowledge hundreds of teachers who participated actively in these researches and we are grateful to them.

CONTENTS

ON THE RELIABILITY OF RATINGS IN INTERVIEWS

It is almost universal practice to use 'interviewing technique' for purposes of selection of individuals for jobs, positions, admission to further courses etc. Its form has varied over the years but essentially the technique has retained its place. Some 'established' schools even conduct interviews to admit children to kindergarten classes. Selection procedures as opposed to methods of achievement measure must be used to pick up or select the right kind of persons for jobs, positions, scholarships and admission to further courses. Here the main objective will be to find methods, forms and modes that will distinguish and differentiate the person (from others) who will fit into the specifications of job requirement, position requirement or traits required to be successful in the future course of studies. Very often the traditional and more established procedures at all levels have a tendency to eliminate rather than select just the same way our traditional examinations try to find what students do not know instead of trying to assess what they know. The aim of this paper is in part to look at some forms while the major emphasis will be given to present a method of assessing the reliability of 'interview' ratings and suggest ways and means of introducing validity and meaning to our selection procedures with interviewing techniques as one of the components.

Some of the methods where interviewing technique is used for selection of persons for jobs, positions and admission to further courses are discussed below.

I. Interview by a Panel: This is the usual and traditional method. When applied to 'selection' of persons for jobs and positions, applications (giving bio data, qualifications, experiences supported by certificates and testimonials) will be scrutinised carefully by a panel and those who satisfy the assumed criteria will be called for an 'interview' with a panel of selection committee members. A few of them will be from 'outside' and called 'experts' while a few will be those concerned with the organisation, the Head of the organisation and the Head of a Division where the job is positioned. Very often the panel members would be meeting for the first time just a little while before the intended interview: very often they would not have had enough information about the job and its requirements. Very often again they would not have prepared questions to be posed. There has been so much of adhocism in the whole process of 'interviewing' and fairness of selection 'questionable'. All kinds of 'malpractices' can enter and have entered in the past with the result everyone has lost faith in this system. While this is a form of selection that will certainly live with us for quite some time, it is disheartening to find that attempts at improvement of this (with a view to make it more and more scientific) have been few and far between. Certain suggestions are given here to improve this system.

(1) The job and its description will be given to every aspirant.

(2) Application blanks will have to be scientifically prepared soliciting all kinds of relevant data and information about the applicant in respect of qualifications and experience and personality characteristics like hard work, initiative, originality, attitude, etc.

(3) Data and information collected will have to be processed scientifically with a suitable system of weightages and on the basis of this data processing, applicants will have to be listed in order of merit to be called for interview.

(4) Wherever possible, those who have been selected may be asked to take a selection test (fixed and free response types) which is prepared to fit very closely to specifications of the job requirement.

(5) An interview trying to assess the person's ability to express argue, ability to think clearly, ability to judge and evaluate, etc., may be conducted. The panel with experts must prepare well in advance certain questions some involving fixed response and others free response for the interview, The author has used with advantage a procedure where questions are written in small pieces of paper and given to interviewees for answering.

II. Interview by a Panel sitting individually: Very often for purposes of admission to further courses (professional courses especially), interview technique involving a panel sitting individually and judging the performance is what is adopted. Certain criteria are usually laid down and weightages allotted for achievement in the previous qualifying examination and the performance in the interview. It is possible in this procedure to take care of various factors and this lends itself quite appropriate for evaluation. At the same time there is plenty of scope for malpractices. Certain suggestions can be given to make this procedure much more meaningful and valid.

(1) The panel must meet and sort out heads of assessment and procedures of assessment. They must think of more objective forms of checking the traits they are going to measure.

(2) Proper weightages have to be given and perhaps communication to interviewees before hand.

(3) Ratings (rank order, mark, etc.) will have to be compared and if possible the reliability of these ratings ascertained. It is absolutely necessary to give weightages to written tests, interviews, etc., based on the reliability of these component measures.

III Successive Interviews: There are some organisations (usually the private sector) who select their persons on the basis of a series of interviews, the preliminary ones to find out suitable people who can stand upto their more rigorous interviews later. A system of preliminary interviews and a final interview is in practice. Where this is done, very often the tendency has been to eliminate quite a good number at the preliminary stages and keep a handful for the final interview. This happens whenever (in our country) private sector wants to recruit suitable persons for their jobs. Regional interviews (preliminary) will be held and at every centre in the region, a few (two or three usually) will be selected for a final interview at a central place. Usually the final interview will be in the form of a written test together with interview with a panel all sitting together. Recruitment to the services (Army, Navy and Air Force) is also of similar pattern.

IV. Selection Based on outstanding Performance in the Previous Qualifying Examination: This is also in practice in firms (particularly Engineering Industry, Pharmaceutical, Heavy and Medium Industries) to offer a kind of apprenticeship training or officer training to those who have shown outstanding performance in the

previous qualifying examination. Very often the training given to them is so comprehensive and exhaustive that they do not recognise the merit of their performance in the previous qualifying examination beyond the point of selection.

V. Interview After a Written Test: Many organisations (Banking, Insurance, Public Sector Project Undertakings) have their own well established and well prepared written tests administered. These tests very often combine objective type, short answer, essay and problem solving questions together with attitude/interest scale and performance questions. Here again certain suggestions can be given to make these written tests still better. Some of these are:

(1) The specifications for these tests (for making questions/items) should be derived from job descriptions and job requirements. A systematic job analysis (combination of questionnaire, interview, on the job work-time studies, etc.) must be done and description and requirements derived on the basis of successful job performers. The author* successfully completed job analysis of technician level jobs in Automobile Engineering for a system design of a 6 semester technician diploma course. These tests will have to be made by a panel of subject experts, job performers, evaluation experts all sitting together and analysing the specifications.

(2) In order to improve reliability (selection type tests must have a very high degree of reliability) most of these must be of objective type and short answer type.

(3) Over the years, the performances of these selected on these tests must be correlated with the performances on the job and if necessary, tests revised on the basis of 'feedback'.

VI. Selection Based on On-the-Job Trials: Even though this is seldom used in Indian context, this has a tremendous potential. Teachers, Salesmen, Managers and the like can be selected on the basis of their performance of teaching to a small group; (observed by raters with a checklist of criteria for evaluation) on the basis of actual selling and on the basis of managing for a specified period of trial time. The author is aware of an Institution that tried to select a Head of a Department by attaching the interviewees one day each to various departments and evaluating their work (This has a very high validity and to make it reliable, every department will be given a check list of criteria with which to judge.

VII. Selection to Jobs/Positions by 'Invitation': Sometimes for top positions and specialised jobs, a panel of outstanding persons' names will be made and they will be 'invited' to take up these positions. Usually Vice-Chancellor for Universities Members of Commissions, Members of Committees will be selected on this basis. The success of these persons in these jobs are purely based on 'chance' factor.

*Natarajan V. 'A system approach to the design of Educational and Training Systems in India' - paper to be presented at an International Conference in Cairo on Nov. 25-28, 1977, in Proceedings by Pergamon Press, London, 77.

After having reviewed these methods involving 'interview' technique of one form or the other, let us take up the most important question of the degree of reliability (agreement, consistency, acceptability) of such an interview procedure and also its relations with the component of written test performance.

The ratings by three selection committee members (independent of each other) of 7 candidates who appeared for an interview are given below:

TABLE

S. No.	Person	Rater I	Rater II	Rater III	S	S^2
1	A	5	2	5	12	144
2	B	1	1	2	4	16
3	C	6	4	7	17	289
4	D	3	3	4	10	100
5	E	2	7	1	10	100
6	F	7	5	6	18	324
7	G	4	6	3	13	169
		28	28	28	84	$114^2=\Sigma S^2$

Here k = 3, no. of raters

N = 7, no. of those rated

We have here the rank positions of the seven persons as given by three raters.

Applying formula $*\bar{p} = 1 - \frac{k(4N+2)}{(k-1)(N-1)} + \frac{12\ \Sigma S^2}{k(k-1)\ N(N^2-1)}$

Where $\bar{p}$ = average inter-correlation among individual judges

k = No. of judges

N = Number of those rated or stimuli

S = Sum of the ranks for any stimulus or person

*Guilford - Psychometric methods - page 253

The average rank order inter-correlation

$$\bar{r}_{11} = 1 - \frac{3(28+2)}{(2)\ (6)} + \frac{12 \times (11.42)}{(3)\ (2)\ (7)\ (48)} = 0.30$$

What will happen if the number of rater is doubled? We can apply Spearman Brown formula.

$$r = \frac{2 \times \bar{r}_{11}}{1 + \bar{r}_{11}} = \frac{2 \times 0.3}{1 + 0.3} = \frac{0.6}{1.3} = 0.46$$

The average rank order inter-correlation or reliability of rating by 6 raters will be 0.46. While those of 3 raters is seen to be 0.30.

The most recently suggested method of estimating reliability for ratings has been described by Ebel* . If each of k raters has rated N persons on some trait on one occasion, we have the possiblity of obtaining inter-correlations of ratings of the N persons from all possible pairs of K raters.

This suggests the use of the statistic known as the intra class correlation, which gives essentially an average intercorrelation. Ebel's formula is:

$$\bar{r}_{i}^{1} = \frac{v_p - v_e}{v_p + (k-1)\, v_e}$$

Where $\bar{r}_{1}^{1}$ = reliability of ratings for a single rater

v_p = variance of persons

v_e = variance for error

k = number of raters.

It should be noted that this gives the mean reliability for one rater. The reliability of the mean of k ratings for each person would be greater. For this Ebel gives the formula.

$$r^{KK} = \frac{v_p - v_e}{v_p}$$

This gives the reliability for mean ratings from k raters. One could arrive at the same result by applying to this reliability for 1 rater, the Spearman Brown formula to predict the reliability for a measure k times as long.

*Ebel R. L. - Estimation of the reliability of ratings, Psychometrika, 1951, 16, 407-424.

Ratings of seven persons made by 3 raters (for a group of the same accepted traits) prepared for determining variances used in estimating reliability of ratings:

S. NO.	Person	Rater I	Rater II	Rater III	X_p^2	ΣX_p^2
1	A	5	8	5	18	324
2	B	9	9	8	26	676
3	C	4	6	3	13	169
4	D	7	7	6	20	400
5	E	8	3	9	20	400
6	F	3	5	4	12	144
7	G	6	4	7	17	289
Σ Xr		42	42	42	126 $=\Sigma X_p$	2402 $=\Sigma(\Sigma X_p)^2$
$\Sigma(Xr)^2$		1764	1764	1764	5292 $=\Sigma(X_r)^2$	

$$\Sigma X^2 = 840 \qquad \frac{(\Sigma X)^2}{KN} = \frac{126^2}{3 \times 7} = 756$$

The sum of squares for persons is

$$\Sigma d_p^2 = \frac{(\Sigma X_p)^2}{k} - \frac{(\Sigma X)^2}{k\,N}$$

$$= \frac{2402}{3} - 756 = 44.66$$

The sum of squares for raters is

$$\Sigma dv^2 = \Sigma \frac{(\Sigma Xv)^2}{N} - \frac{(\Sigma X)^2}{k\ N} = 0$$

The total sum of square is

$$\Sigma X_t^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{kN}$$

$$= 840 - 756 = 84.00$$

And finally, the sum of the squares for remainder or error is

$$\Sigma d_o^2 = \Sigma X_t^2 - \Sigma d_p^2 - \Sigma d_r^2$$

$$= 84. - 44.67 = 39.33$$

Computation of variances needed to estimate Reliability of the ratings

Source	Sum of squares	Degrees of freedom	Variance
From persons	44.67	6	7.445
From raters	0	2	--
From remainder	39.33	12	3.2775

The sums of squares are given above, their degrees of freedom and the two variances we need V_p & V_e

$$f\ \bar{r}_{11} = \frac{7.445 - 3.2775}{7.445 + (3 - 1)\ 3.2775} = 0.298$$

This is the reliability for one rater. For the three raters combined or for the averages of their ratings.

$$r_{33} = \frac{7.445 - 2.775}{7.445} = 0.560$$

Spearman Brown formula for 3 raters from 1 rater of 0.298 gives the same result.

It may be seen that the earlier formula is quite easy and less time consuming. The author recommends the use of this formula in preference to Ebel's formula.

Conclusions

The reliability of ratings for one rater is found to be 0.298 or 0.30 (by less tedious formula). The rating of any individual has a consistency or reliability of

0.30 only. This means that for 30% of cases only, there will be agreement for ratings given by the individual. One conclusion is to take any one rater's ratings and give it only 30% credit. If there is a written test component alongwith the interview, the written test can be given 70% credit and the interview 30% credit. The total is worked out from every candidate and the person who has the highest score can be selected. Another procedure is to average the ratings (combine the three persons' ratings) and give it in our case a credit of say 60% and the written test the balance of credit of 40%. The person who finishes at the top can be selected. It is, therefore, necessary to work out the reliability of ratings of one rater or all the raters put together and take this into account suitably. It is hoped that all agencies like U. P. S. C., Universities, College, Firms, Industries engaged in selecting persons for jobs, positions, admission to further courses, will do similar calculations for reliability of ratings and suitably take this into account thus assuring credibility in their decisions.

In the 'Examination Reform - A Plan of Action' circulated to all universities in the country, the UGC observed this: "A large body of teachers and educational administrators is not yet fully conscious of the subjectivity, unreliability and lack of validity of the examinations as conducted today". Talking of unreliability, we must clearly understand that the mark given by an examiner is something like a 'raw mark' and it is certainly different from his 'true mark'. Such a raw mark is subject to an error which for some typical papers set at universities (the UGC Report continues) is greater than 5 per cent. This means that when an examiner assigns a mark of 43 the true mark may be either above 48 or below 38 in 50 per cent of the cases. There are many sources of error that contribute to unreliability of our present day examinations. The author* has drawn attention to this and lists the following: a) Error due to subjectivity in marking, b) Error due to biased sampling of topics, c) Error due to biased sampling of abilities, d) Error due to allowing students a choice of questions, e) Error due to arbitrary time limits, f) Error due to assumptions in addition of marks. Concluding this section, the author remarks; 'If universities/exam. boards in our country at all levels take steps to indicate along with raw marks in each subject for a student details like mean, standard deviation, reliability coefficient and error in measurement, there will then be no need for any treatment of scores.

It is possible to look at very closely one of the sources of error in traditional essay type examinations that due to allowing students, a choice of questions the presence of question choice in public university examinations of the traditional (essay) kind has been accepted and indeed upheld for many years in India at the University level. The reasons given for the need to allow this choice of questions are many: but two of the most important ones are: (i) it allows the teachers freedom to teach particular portions of the syllabus (in which they may be particularly interested), (2) it allows students to concentrate on particular aspects of topics in which they are able to show themselves to the best of advantage. This has led to the undesirable situation that teachers indulged in dealing with only a few topics in the syllabus leaving the rest for choice and on the other hand students increasingly indulged in 'selective cramming'. Both these have serious implications on teaching-learning, impairing its effectiveness, efficiency, relevance adequacy and above all purpose. Question choice is still accepted today alongside the adoption of more rigorous methods of examining (e.g. the use of objective items) but little thought seems to have been given to the problems raised by allowing a choice of questions or even recognizing the fact that complications may occur. Some of these problems are presented here. Before presenting the problems, it is as well to remember the basic assumption that is made (albeit implicitly not explicitly) when using an examination where a choice of questions is presented. This assumption is that a candidate will be able to be compared with others taking the same examination whatever combination of questions he attempts on an exam paper. This implies a form of 'currency' (or comparability between individual questions) and hence combinations of questions. The viability and validity of this assumption will be explored now.

*Natarajan V. — 'Monograph on Grading for Universities; AIU 1976-pp 35-37.

The Problems and their Effects

(A) The Syllabus: It may be asked whether or not syllabus topics (in their own right) are of the same basic level of difficulty. In general, it is unlikely they are, but such an assumption one way or other could be made only on the basis of a consensus of opinion. Let us consider a few examples. In undergraduate mathematics, a quoting of a principle in complex variable and an example by students and a factorization followed by the solution of a pair of simultaneous equations cannot be considered 'equal' in basic level of difficulty. Here we are considering only the content and nothing else. In Geography again, is the description of industrial growth of an area on a par with naming rivers, hills and features on a map? We are now considering basic complex variable as opposed to solution of simultaneous equation and also a knowledge of a terrain. It is likely that no agreement is reached regarding an answer to this problem. This is not important but what is important is that as long as this difference of opinion in regard to the 'equality' of subject topics exists, then the 'equality' of the results of candidates attempting these questions may be questioned.

(B). Abilities: Let us look at these questions in terms of the 'abilities' they purport to test. Some of them may clearly memory and only memory. Yet others do involve 'comprehension' while some others deal with 'application' ability being put to test. How far are we justified to put on par a pair of questions that is definitely known to be testing different 'abilities'? Some questions will involve the students to translate information given in one form into another (from verbal to graph). Yet some other questions will involve calculations (after recalling a set of formulae, rules, procedures) and infer from these calculations granting that the questions test what they intend to test, then these questions in a paper will be testing different 'abilities', and we are not justified to permit 'choice' of these questions. It is necessary therefore to group questions of the same category in terms of equality of ability tested with sections and get students to respond to these sections. The restructured pattern* proposed by AIU (and accepted by many universities) involves in every paper:

Part A: Objective type 20 to 40 items/20 to 40 mins/20 to 40 marks.

Part B: Short answer type 10 to 15 questions/50 to 90 mins/40 to 60 marks.

Part C: Long answer essay type 1 to 3 questions/20 to 60 mins/15 to 30 marks.

Part A may have items (M/C, MF, etc.) all of them testing knowledge/comprehension. Part B may have questions, all of them testing Analysis/ Application and Part C may have questions all of them testing synthesis/ evaluation. In this, it may be noted that there is no choice in Part A/ Part B and there may be internal choice in Part C.

(C) The Difficulty of Questions: It has been the practice to allow a choice of questions only in respect of supply type questions whether SA or LA. Where a restructured pattern is in practice, only LA questions give a choice of questions for

* 'A Restructured Pattern of Exams for Universities', Natarajan V, University News, AIU, Delhi.

students. In respect of long answer questions, some of them are inherently more difficult than others. Elsewhere* the author has shown that it is possible to work out a very realistic and accurate difficulty index (or Facility value) for every question in a choice type exam. In fact there are two indices.

F. V. index = 50 + (M_Q - M_T) where M_Q = mean percentage mark on the question by those attempting it and MT = mean percentage total mark on the whole test by those attempting it. This is found to be a 'sample free' technique. If we analyse our choice type exam on this basis, we will come up very quickly that our questions range considerably in F. V. One such analysis is given below:

Choice type exam. with question No. II compulsory and any 5 to be answered in 7 questions. Totally there are 8 questions. Question No. I carries 15 marks and question No. II to question No. VIII - 10 marks each.

QN No.	I	II	III	IV	V	VI	VII	VIII
M_Q	77.3	71.5	72.8	51.9	15.0	57.5	31.9	61.7
M_T	63.13	66.0	63.7	74.3	52.7	63.4	58.8	63.13
$(50+M_Q-M_T)$ = F.V.	64.0	55.5	59.1	27.6	12.3	44.1	23.1	48.57

It has been argued that the difficulty of a question (or F. V) is a function of both different ability groups attempting different questions and also leniency and severity of marking. Another important thing we are concerned with here is that supposing question III is taken in the above situation (forming a part of the choice exam) and its Facility Value as such compared with its F. V. of the question were to appear in a No Choice Exam, we find that different values result. Certainly, 'choice' or 'no choice', different questions have Facility Values. To imagine that they are all equal and to give a choice of questions to students is extremely misleading and unjustified to say the least.

(D) Validity: In achievement examinations, we are principally concerned with content validity. It is easily appreciated that 10 questions or so will not be able to cover 100% of the syllabus. Very often the coverage will not exceed 80% of topics. In our traditional papers, the usual practice is to give 10 questions and ask students to

* Natarajan V - 'Monograph on Test & Item Analysis for Universities, AIU, New Delhi.

answer 5 out of these. Immediately the validity is reduced to 40%. If we set pass mark in this paper as 35 out of 100, it really means that a person who has mastered just 0.35x40 = 14% of syllabus is declared passed. In constrast to this situation, we consider our restructured pattern:

Part A	20 to 40 items	No Choice
Part B	10 to 15 questions	

and part C 1 to 3 questions with internal choice, if any, then it can be seen that at the first place, the content validity of the paper is considerably increased say 90 to 95% and almost all questions are compulsory. Thereby, the validity remains more or less 90% It is also usual practice to increase the minimum marks for a pass to 40 or very often 50 in such a restructured paper. The overall validity for passing comes out to 45% as opposed to 14%.

(E) Reliability: When we consider the reliability of a choice type examination compared to that of a no choice type, may things can be said. The overall reliability or the index of measurement efficiency is very high in a no choice type, since the number of items are more and the ratio of error variance to observed variance is very small. The internal consistency reliability of a no choice is also more than that of a corresponding choice type examination. However, it must be said that it is much more complicated to calculate reliability of a choice type compared to no choice type

(F) Uniformity: 5 questions out of 10 can be chosen in $10C_5 = 252$ different ways. There is in fact 252 different combinations in different papers. Performances by students in this situation can never be compared. There is uniform injustice shown to students.

Some of the problems (and their effects) relevant to allowing a choice of questions on examination papers have been presented here. At the same time it has been demonstrated that as far as all the problems are concerned, different rules will apply to different candidates since they all attempt different combinations of questions. The syllabus content, the ability, facility value are all bound to affect the marks of those taking it; in the situation where choice is presented, however, depending on the combinations, these factors will have an (unknown differential) effect. It is this unknown effect working in a different manner from one student to another that causes the greatest uncertainty in accepting the results from a choice type exam.

From time to time, certain ad-hoc solutions have been suggested for this problem. The earliest one to be adopted was to have the paper divided into sections and choice allowed within sections. Another way is to have a question or two as compulsory and a marginal choice (5 out of 7 or 4 out of 6) given. Another method tried was to keep certain questions starred and others free. Students may be asked to attempt starred questions for at least 2/3rds maximum marks and the rest only from

unstarred. All these 'ad-hoc' solutions have a serious limitation. They have all involved traditional type (essay) and a fewer number of questions. A rational and scientific way is to restructure the examination as discussed early in this paper. This alone would bring in improved validity, reliability, relevance and meaning into the process of evaluation of student performance. To implement this restructured pattern, the Chairman, Boards of Studies in different subjects together with his members must be given the orientation and training to enable them to take decisions on patterns. However, it must be said that within the rigid framework, there is flexibility. Rigidity is to be understood in terms of Part A -objective type, Part B - Short answer type and C - Long answer/problem solving being common for all subject areas but flexibility is there for different Boards in the decision of the number of items, duration of parts and marks allotted to parts. A pattern for different subjects at undergraduate and at Postgraduate levels adopted in one university was sent to many universities to suggest modifications and the consensus arrived published in the form of an aritcle*

In a typical subject the pattern may be:

Part A	:	Objective type (Mostly M/C and M/F) 30 items for 30 minutes and 30 marks.
Part B	:	12 S. A. questions for 90 minutes and 50 marks
Part C	:	2 L. A. questions for 60 minutes and 20 marks.

In this pattern, the total number of questions to be answered by a student is 44. With 44 questions, the content validity of the paper is very high. There is objectivity in marking Part A and Part B. There is however an element of subjectivity in Part C marking. At the same time, the overall reliability of the whole exam is very high. There is relevance; there is meaningfulness, there is purpose above all better validity and reliability. In addition to all these such decisions about the pattern by Boards of Studies will have far reaching implications - indeed this decision of pattern will influence the nature and kind of Question Banking, the relative weightages of internal and external and above all science and rationality in our evaluation of student performance.

* Natarajan V - "A Restructured pattern of exams for Universities", University News, AIU.

The main purpose of this paper is to highlight on the simple statistical methods that yield fairly acceptable and reasonably accurate results which may be of great help to those who are not mathematically inclined to appreciate sophisticated methods. It is also felt that formats used in this section will help the teachers feel confident to undertake test and item analysis of all their class tests. These simple methods will be illustrated with the example of:

a) a 20 item objective type test on 76 students and

b) on actual choice type drawing examination of 8 questions on 117 students.

SIMPLIFIED METHODS IN TEST ANALYSIS:

a) Mean, Median and Mode: Mean of scores is the usual addition of all scores divided by the number of students. For all these three quantities there are no short cut methods. Median is the middle student's score while mode is the often repeated score. It is possible however for one to get an idea of the value of Mear of scores roughly by looking into the highest and lowest marks and averaging them out.

In our example 20 item test on 76 students: highest was 18 and lowest was 6; the average is $\frac{18 + 6}{2} = 12$ which is nearly the Mean of scores = 12.9 ($\approx$13)

b) Standard Deviation

i) For objective type test: S.D.

$$= \frac{\text{(Sum of 1/6th highest scores - Sum of 1/6th Lowest scores)}}{\text{1/2 total number of students}}$$

This formula is suggested by Jenkins of Lehigh University and quoted by Paul Dieterich. Here 1/6th of the total number of students are considered (in our example 1/6 x 76 = 12.67 $\approx$13 students).

Sum of top 13 student's scores = 220

Sum of bottom 13 student's score = 113

$$\therefore \text{ S.D} = \frac{220 - 113}{38} = 2.81$$

We can take scores of 12.67 student's also (top)

= (18 x 4 + 17 x 4 + 16 x 4.67)

= 72 + 68 + 74.72 = 214.72

We can take scores of 12.67 student's also (bottom)

$$= (6 \times 1 + 7 \times 1 + 8 \times 2 + 9 \times 6 + 2.67 \times 10) = 109.7$$

$$\therefore \quad = \frac{214.72 - 109.7}{38} = \frac{105.02}{38} = 2.77$$

Actual standard deviation calculated on the basis of the RMS Value of deviations of individual scores over the mean = 2.79.

ii) For choice type examination: There are two alternatives:- One is to use Jenkins formula.

$$\text{S.d} = \frac{\text{(Sum of 1/6th highest - Sum of 1/6th lowest)}}{\text{1/2 x total number of students}}$$

the other is to use Harper's formula

$$\text{s.d} = \frac{\text{(Sum of 1/5th highest - Sum of 1/5th lowest)}}{\frac{1}{2} \text{ x total number of students}}$$

1/6 x 117 = 19.5 $\approx$ 20 students

1/5 x 117 = 23.4 $\approx$ 23 students

Sum of top 20 student's scores = 1563

Sum of bottom 20 student's scores = 1062

$$\text{S.d (Jenkins)} = \frac{1563 - 1062}{58.5} = 8.55$$

Sum of top 23 student's scores = 1781

Sum of bottom 23 student's scores = 1239

$$\text{S.d (Harpers)} = \frac{(1781 - 1229)}{58.5} = 9.25$$

The actual value calculated = 8.34. So we can say that Jenkins formula is a better one.

iii) Standard Error of the Mean $= \frac{\text{s.d}}{\sqrt{N - 1}}$

Use of the approximate value of S.D.

iv) Index of Measurement Efficiency

$$IME = 1 - \frac{0.16K}{\sigma^2} \quad \text{(McMorris)}$$

Where K = number of items

σ^2 = Variance of test scores = $(sd)^2$

This can be used only for objective type tests. The smaller the number (k), the lower the estimate of the index and any value above 0.75 is likely to indicate a very good reliability for the test. For large values of k, this formula will yield comparable results with other methods.

Example 20 item test on 76 students

$$IME = 1 - \frac{0.16 \times 20}{\sigma^2} = 1 - \frac{3.20}{7.79} = \frac{4.59}{7.79} = 0.59$$

This is almost the same as reliability by many methods. For greater values of K there will be remarkable identity of values between this McMorris formula and other reliability estimates.

v) Reliability:

a) For objective type test:

1) Split half reliability between two measures x and y if correlation is required (Product Moment), we have seen that

$$r = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{N}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{N}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{N}}}$$

If the class size is small (upto 40) odd-even score correlation that gives split halves reliability can be calculated by this.

A simple formula is to use $r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N S_x S_y}$

Where $\bar{x}$ = Mean S_x = s.d. of x

$\bar{y}$ = Mean S_y = s.d. of y

N = Number of students

Yet another simple method for finding correlation between two measures (odd, even scores etc.), tetrachoric 'r' can be read from the following table

%	r	%	r	%	r	%	r	%	r
45	0.95	37	0.69	29	0.25	21	-.25	13	-.69
44	0.93	36	0.65	28	0.19	20	-.31	12	-.73
43	0.91	35	0.60	27	0.13	19	-.37	11	-.77
42	0.88	34	0.55	26	0.07	18	-.43	10	-.81
41	0.85	33	0.49	25	0.00	17	-.49	9	-.85
40	0.81	32	0.43	24	-0.07	16	-.55	8	-.88
39	0.77	31	0.37	23	-0.13	15	-.60	7	-.91
38	0.73	30	0.31	22	-0.19	14	-.65	6	-.93

To use the table : Find the percentage of students who stood in the top half of the group on both measures you are correlating and look up the correlation (r) corresponding to this percentage in the above table.

These are called tetrachoric correlations while the more common but more difficult kind are called "Product-Moment Correlations". They mean the same thing, in the sense that the tetrachoric yield a fairly accurate estimate of the correlation that you would get by the product moment method. Tetrachoric r's are perfectly respectable and are often used in educational research but they are not very precise since a difference of 1% can make a difference as great as 0.07 in the correlation. However the reliability of the data that teachers usually have to work with and the relatively small number of students involved do not justify more precise methods of computation. A rough idea of relationship is enough.

Since even 1% of the students can make so much difference in the correlation it is important to use, a standard, uniform method of counting how many students stood in the top half on each measure. To find the middle score on each measure:

I. List the scores on each measure from highest to lowest and put a tally after each score for each student who made it.

II. After all the scores have been tallied, count down the tallies to half the number of students in the group.

III. The score at which this middle tally falls is the middle score.

You will ordinarily have the students listed in alphabetical order and after each name, you will have the two scores that you are correlating. After you have found the

middle score on each measure, go down the list and put a check (✓) after each score that stands above the middle score on that measure; a straight line after each score that stands at the middle score. Do this separately for each of the two measures?

Then if you need six students with middle scores can Measure A to take in half of the group, put a check (✓) through the first six straight lines on Measure A that you come to in alphabetical order. If you need four more students with middle score on Measure B, put a check (✓) through the first four straight lines after the scores on that measure. The count how many students have two checks (✓) after their names; turn this number of students into percentage (not by the number in the top half) look up this percentage in the previous table. The decimal corresponding to it will be the correlation between the two measures.

It is not necessary for the two measures to be on anything like the same scale. It is perfectly valid for example to correlate the height in inches with weight in pounds; or scores on an objective test that run from 200 to 800 with scores on an essay that run from 1 to 9. All that is necessary is to count how many students stood in the top half of this same group on both measures.

It is impossible and meaningless, however, to correlate the scores of two different groups on the same measure, for example, to correlate scores of boys with those of the girls teachers often speak loosely of correlating one class with another when they really mean comparing. There is no way to correlate two groups of students on the same measure; one can only correlate two sets of measures on the same students. To compare the performance of two groups of students on the same test or measure, you compare their averages and if you want to find out whether the average is 'really' different you compute the standard errors of these averages and then the standard error of the difference.

The topic of correlation is related to the topic of reliability because often the only way of computing the reliability of a test is to give two tests of the same ability and correlate the two sets of scores. This is true of essay tests and tests in which the items receive different number of points. The KR formula is applicable to objective type tests.

Rulon Formula (odd-even) $r_{tt} = 1 - \frac{\sigma^2_d}{\sigma^2_t}$

Where d = difference between two half scores of an examinee.

d = S.D. of those differences

t = S.D. of total scores

S.D. of differences is only to be found

K.R. formula 20 : $\frac{n}{n-1} \quad 1 - \frac{\sum_{i=1}^{n} p_i q_i}{\sigma_o^2}$

Individual question variances can be found out.

Approximate S.D. calculation formula can be used

Choice type Nuttal (1969) N_4 formula can be easily applied.

$$r = \frac{1}{1-1}\left[1 - 1\left(\sum_{i}^{k} n_j \; S_j\right)^2 \Big/ S_x^2 \sum_{i}^{k} n_j\right]$$

1 = number of questions to be attempted

K = number of questions set

n_j = number of students who answered question j

S_j^2 = variance of scores on question j

S_x^2 = variance of total scores

Standard Error of Measurement

$$SEM = Sd\sqrt{1 - r}$$

where Sd = value calculated by approximate methods

r = least reliability coefficient

This will give the class teacher an idea of the Standard Error of Measurement.

For Objective Type Tests

$$S.E.M = \sqrt{0.16\,k}$$

Where K = number of items

It is an approximation derived from McMorris formula.

$$IME = \left(1 - \frac{0.16\,k}{\sigma^2}\right)$$ Since 0.16 K account for "error variance".

For Choice Type/Essay Type

$$SEM = Sd\sqrt{1 - r}$$

Derived Scores

The derived scores that can be worked out by the classroom teacher are Z scores T scores, CEEB and percentile rank.

$$Z \text{ scores for every individual} = \frac{(\text{Raw score} - \bar{x})}{s.d} = (Zsay)$$

($\bar{X}$ = Mean)

T score for the individual = 10 Z + 50 (T say)

CEEB score for the individual = 10 T

SIMPLIFIED METHODS OF ITEM ANALYSIS:

Item analysis or question analysis (as applied to supply type) consists in analysis for two important characteristics; facility value and discrimination index of every item/question. While there are sophisticated and very rational techniques available, certain simplified methods are given here to help teachers/researchers (who are not quite competent in statistics to calculate these characteristics.

a. Facility Value

1. For objective type test

$$\text{F.V. of an item} = \frac{\text{Number answering the item right}}{\text{Total number attempting the item}} \times 100$$

1) Total population can be taken into account when the class strength is small (< 40).

2) When it is between 40 and 100, top-bottom 27% alone needs to be considered

3) When class strength is beyond 100, top-bottom 10% will be sufficient.

2. For Choice Type Questions

There are two alternatives

i) F.V. of a question = $(50 + M_Q - M_T)$ (Morrison)

Where M_Q = Mean percentage mark on that question for all candidates.

M_T = Mean ability in percentage total mark of those who attempted the question.

a) if class strength is upto 40, entire population can be taken into account.

b) if strength is between 40 and 100, top-bottom 27% alone needs to be considered.

c) if strength is $>$ 100, 10% top-bottom will be sufficient.

ii) F.V. of a question = $M + M_Q - M_T$

Where M = Mean % marks (Willmott, Nuttal) Dr. Edwin Harper, however has suggested the following formula for F.V.

$$F.V. = \frac{M_u + M_L}{2 \times \text{Max. Marks for the question}} \times 100$$

M_u = Mean mark of top 1/5th of students on the question

M_L = Mean mark of bottom 1/5th of students on the question.

b) Discrimination Index

1. For objective Type Test Items

D.I. of an item = (F.V.) higher ability 27% - (F.V.) lower ability 27% (Johnson upper lower index). This is quite in order for classroom tests.

a. if strength is $<$ 40, 50 - 50 grouping can be done (or 27% - 27% also).

b. if strength is between 40-100, 27% - 27% will be all right.

c. if strength is $>$100, 10% grouping will be in order

2. For Choice Type Questions

There are two alternatives:

a. D.I. is the product moment correlation between marks on the question and total marks.

This can be done in the usual way (or) it shall be easily calculated from the formula:

$$D.I. = \frac{\Sigma (x - \bar{x})\ (v - \bar{y})}{N\ S_x\ \ S_y}$$

Where x = marks on question (whose DI is to be found) of a given students.

y = total marks of the same students

N = Number of students

S_x = S.d. of X

S_y = S.d. of y.

b. D.I. is the product moment correlation between marks on the question with (Total marks - that question marks) No simple method is however possible if we retain the basic definition of D.I. Dr. Edwin Harper has suggested a simple formula for D.I.

$$D.I. = \frac{1.8\ (M_u - M_l)}{\text{Max. marks for the question}}$$

5. Derived Scores For Every Student

$$Z = \frac{\text{Raw score} - \bar{x}}{\text{S.D.}} = (Z \text{ say})$$

$$T = 50 + 10\ Z$$

$$CEEB = 10T$$

Objective Type Test Analysis

Test Number Time

Number of Items Date

N(Number of answer sheets 1/6 N =

Upper group/Lower group 27% of N =

K = number of test items 0.16 k

C = average number of choice/item

1. Mean $\bar{x}$ = $\frac{\text{Sum of all scores}}{N}$ =

Median = Middle student's score =

Mode = Frequent score =

2. **Standard Deviation** = $\dfrac{\text{(Sum of 1/6 top)} - \text{(Sum of 1/6 bottom)}}{N/2}$

When N is small use (N-1) instead of N in the denominator)

3. **Reliability/Index of Measurement Efficiency**

$$= r = \left(1 - \frac{0.16\,k}{S.D.^2}\right)$$

4. **Standard Error of Measurement**

$$S.E.M. = \sqrt{0.16\,k} =$$

$$= S.D.\sqrt{1 - r}$$

STATISTICAL ANALYSIS SHEET

Course Subject Total Scripts_____

Date Number in High Group

Number in Low Group

d	1-d	d(1-d)	Column X ITEM	a N_H%	b N_L%	c $\frac{a+b}{200}$	d $\frac{a+b}{100}$	e d(1-d)	d	d	d(1-d)
0.99	0.01	0.01	1	90	25	0.58	0.65	0.23	.74	.26	.19
0.98	.02	.02	2	95	70	0.83	0.25	0.19	.73	.27	.20
0.97	.03	.03	3	95	55	0.75	0.40	0.24	.72	.28	.20
0.96	.04	.04	4	90	60	0.75	0.30	0.21	.71	.29	.20
0.95	.05	.05	5	75	15	0.45	0.60	0.24	.70	.30	.21
0.94	.06	.06	6	100	60	0.80	0.40	0.24	.69	.31	.21
0.93	.07	.07	7	95	65	0.80	0.30	0.21	.68	.32	.22
0.92	.08	.07	8	90	30	0.60	0.60	0.24	.67	.33	.22
0.91	.09	.08	9	70	40	0.55	0.30	0.21	.66	.34	.22
0.90	.10	.09	10	80	15	0.48	0.65	0.23	.65	.35	.23
0.89	.11	.10	11	70	20	0.45	0.50	0.25	.64	.36	.23
0.88	.12	.11	12	85	30	0.58	0.55	0.25	.63	.37	.23
0.87	.13	.11	13	95	70	0.83	0.25	0.19	.62	.38	.24
0.86	.14	.12	14	60	35	0.48	0.55	0.25	.61	.39	.24
0.85	.15	.13	15	95	65	0.80	0.30	0.21	.60	.40	.24
0.84	.16	.13	16	80	55	0.68	0.25	0.19	.59	.41	.24
0.83	.17	.14	17	90	85	0.88	0.05	0.05	.58	.42	.24
0.82	.18	.15	18	75	75	0.75	0.00	0.00	.57	.43	.24
0.81	.19	.15	19	80	55	0.68	0.25	0.19	.56	.44	.24
0.80	.20	.16	20	15	0	0.08	0.15	0.13	.55	.45	.25
0.79	.21	.17							.54	.46	.25
0.78	.22	.17							.53	.47	.25
0.77	.23	.18							.52	.48	.25
0.76	.24	.18							.51	.49	.25
0.75	.25	.19							.50	.50	.25

ANALYSIS OF CHOICE TYPE EXAMINATION

		YES	NO
NUMBER OF QUESTION	FREE CHOICE	/___/	/___/
NUMBER OF QUESTIONS TO BE ANSWERED	INTERNAL CHOICE	/___/	/___/
	1. COMPULSORY QUESTION & FREE CHOICE	/___/	/___/
NUMBER OF COMBINATIONS	TIME		
NUMBER OF STUDENTS	DATE		
	MAX. MARKS		

PERFORMANCE MATRIX

S. NO.	ROLL NUMBER	QN I	QN II	QN III	QN X	TOTAL

ANALYSIS

1. MEAN = $\frac{\text{Sum total marks of all students}}{\text{Number of students}}$ =

   MEDIAN = MIDDLE STUDENTS SCORE =

   MODE = OFTEN REPEATED SCORE/S =

2. STANDARD DEVIATION : ACCURATE VALUE = $\sqrt{\frac{(x - \bar{x})^2}{N}}$

   APPROXIMATE VALUE = $\frac{\text{(Sum/16th highest-Sum 1/6th Lowest)}}{1/2 \quad \text{Total number of students}}$

3. STANDARD ERROR OF THE MEAN = $\frac{s.d}{\sqrt{N - 1}}$

4. RELIABILITY

   a. NUTALL N4 FORMULA

5. STANDARD ERROR OF MEASUREMENT = SEM = S.d. $\sqrt{1-r}$

6. DERIVED SCORES Z, T, CEEB, P.R.

For further reading the readers are referred to look into the Monograph on Test and Item Analysis for Universities, Association of Indian Universities Rouse Avenue New Delhi - 110002.

Reliability can be defined as the degree of consistency between two measures of the same thing. This is neither a theoretical nor an operational definition but is more a conceptual definition. Considering physical measurements, reliability means consistency of a measure or agreement of a number of measures of the same thing. Usually physical quantities measured (like length, mass, time etc.) are quite stable; instruments can be so chosen (from amongst many available) which will be precise and accurate all leading to measurement of high reliability. On the other hand if a person's level of achievement is measured, achievement itself is unstable; measuring devices such as tests are not very precise and accurate and therefore measurements are not very reliable. Educational measurement is typically much less reliable than physical measurement. Test is a systematic measure for comparing the behaviour of two or more individuals (Cronbach 1960). A test measures the attainment or achievement of a certain trait or objective and usually generates numbers which we use to give a meaning to the variation, in attainment. The variation is both inter individual as well intra individual. With the help of inter individual variability data, it is possible to estimate the intra individual variability. There are many different procedures for estimating the consistency or reliability of measurement. Each procedure allows a slightly different source of variation (error). Differences in scores between individual students indicate differences in ability. An individual's test score may vary. The reasons are; the amount of characteristic being measured may change across time (trait instability); the particular questions asked could affect his score (sampling error); any change in directions, timing or amount of rapport with the test administrator could cause score variability (administrator error); in accuracies in scoring a paper will affect the scores (scoring error) and finally such things as health, motivation, degree of fatigue of the person and good or bad luck in guessing could cause score variability. The variation in a person's score is typically called error variance and the sources of variation are known as 'Sources of Error'. The various sources of error may be expected in an examination are as follows:

1. Error due to subjectivity in marking (Mark - Remark Error)
2. Error due to biased sampling of topics
3. Error due to biased sampling of abilities
4. Error due to a choice of questions
5. Error due to arbitrary time limits
6. Error due to examination conditions
7. Error due to assumption in addition of marks.

Reliability Theory

The theory of reliability can best be explained by starting with observed scores (x). We can think of each 'observed score' as being made up of a 'true score' and an

'error score' such that:

$$X = T + E$$

Where X = Observed score

T = True score

E = Error score

The 'true score' is similar to what is refered to as 'Universe score' (Cronbach et al 1972). It is just that portion of score not affected by random error. Any systematic error (such as a scale always weighing everyone 2 kgs too heavy) does not affect reliability or consistency and so, in reliability theory, it is considered as part of the true stable or unchanging part' of a person's observed score.

Individuals, of course, differ from each other with regard to both their true scores and their observed scores since the errors are assumed to be random, theoretically the negative and positive errors will cancel each other, and the mean error will be zero. Also, if the errors are random, they will not correlate with the scores or with each other. By making these assumptions, we can write the variance of a test as

$$V_o = V_t + V_E \quad \text{——— (1)}$$

Where V_o = Variance of a group of individuals' observed scores

V_t = Variance of a group of individuals true scores

V_E = Error variance in a group of individuals' scores

There are two fundamental questions with respect to V_t & Vo

a) Is V_t = Vo (Is the test reliable?)

b) Does V_t Correlate with Vo (Is the test valid?)

In practice, Vo will exceed V_t. Some of the observed variance (Variation) in test scores will be due to Error. Theoretically reliability is defined as the ratio of true score and observed score variances.

$$r_{xx} = \frac{V_t}{Vo}$$

Equation (1) can be written as : $1 - \frac{V_t}{Vo} \Big/ \frac{V_E}{Vo}$

$$\text{i.e.} \quad r_{xx} = \left(1 - \frac{V_E}{Vo}\right) \text{——(2)}$$

It follows that:

1. Reliability coefficient is always less than 1
2. If all the observed variation is due to error, the reliability is zero. A typical value for 'r' is between 0.4 and 0.6 i.e. nearly half the observed variation in test scores will not be due to variation in student attainment.

The effect of error Variance V_E (low reliability) is to introduce an uncertainty (standard error) into any mark allotment. If this inability of a test (low reliability) to make fine discriminations is ignored, use of observed scores in classifying students will become a spurious activity. It is better to accept that most test forms are crude instruments suitable only for course discriminations. In order to improve the reliability of any achievement test, it is necessary to bring down error variance to the extent possible (1) and (2) above are basic formulae from which most of the commonly written expressions concerning reliability and the standard error of measurement are derived.

Standard Error of Measurement of Observed Scores

It can be seen that reliability increases as error variance decreases, if Vo remains constant or if error variance remains constant and we increase Vo, reliability also increases.

$$Se = SEM = S.D.\sqrt{1 - r_{xx}}$$

Where Se = Standard error of measurement. This is the measure of intra individual variability. Since we often can not test a person repeatedly, this statistic is typically estimated from group data. Theoretically the true score of an individual does not vary. If we retested the same person many times, there would be some inconsistency (error) and therefore the observed scores (X) of this single person would vary sometimes more and sometimes less. Making the assumption that the errors within a person's score across testing sessions are random, the positive and negative errors will cancel each other and the mean error will be zero. Thus the mean of the observed scores over repeated testings in the individual's true score ($\bar{X}_i = T$). It is assumed that these observed scores will fall in a normal distribution about the true score.

$$Vo = V_t + V_e$$

If we think of these values as being obtained from the data for a single individual over many testings, then the true score does not change and hence

$$V_t = 0$$

$V_{oi} = 0 + V_e$ where V_{oi} = Variance of a person's observed scores over repeated testings

This holds only for the case where V_{oi} represents the variance of a person's observed scores over repeated testing. If a test has any reliability at all, V_e will be smaller than Vo for a group of individuals, each tested once, because as a group their true scores will vary, even though for each individual $V_t = 0$

Reliability Estimates

How do we obtain estimates of the theoretically defined reliability? Given one set of observed scores for a group of students we can obtain Vo. Then for equation $r_{xx} = 1 - \frac{V_E}{V_o}$; One must get an estimate of either r_{xx} or V_e in order to solve the equation. Ordinarily one estimates r_{xx} first and then uses equation SEM = $Sd\sqrt{1 - r_{xx}}$ to estimate SEM. There are many methods to estimate reliability but the very common ones are given below.

1. Measures of stability

A measure of stability often called a Test-Retest estimate of reliability is obtained by administering a test to a group of individuals, readministering the same test to the same individuals at a later date and correlating the two sets of scores. The practice effect, length of interval are factors that are to be considered.

2. Measures of Equivalence

In contrast to test-retest procedure, the equivalent forms estimate of reliability is obtained by giving two forms with equal content, means, variances of a test to the same group of individuals on the same day and correlating these results. Equivalent forms of a test are of course, useful for reasons other than estimating reliability. For any curriculum or student evalution procedures, a post test over the same type of material as was done in pretest will have to be given.

These two methods will give quite different values. Which one should be used? It depends upon the purposes of the test. For long range prediction, measure of stability can be used. To make an inference about the knowledge one has in a subject matter area, a coefficient of equivalence will be used.

3. Measures of internal consistency

The above two methods require data from two testing sessions. Very often they are not practicable. It is possible, however, to obtain reliability estimates from only one set of test data. With the exception of split half method, the other estimates are really indices of homogenity of items or the degree to which item responses correlate with the total score.

Split Half

This is theoretically the same as equivalent forms method. Novertheless the split half method is considered a measure of internal consistency because the two equivalent forms are contained within a single test, only one test is administered and it is split into two halves (usually odd numbered, even numbered, first half-second half, random half/the other half), the two subscores are correlated. The correlation coefficient (r1/2 1/2) is an estimate of the reliability of a test only half as long as the original. Spearman Brown formula is:

$$r_{xx} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

Where r_{xx} = estimated reliability of the whole test.

$r_{\frac{1}{2}\frac{1}{2}}$ = reliability of the half test.

The spearman Brown formula assumes that the variances of the two halves are equal. If they are not, this method will give a greater value for reliability than other method of internal consistency.

Kuder-Richardson Estimates

One way to avoid the problems of how to split the test is to use one of the KR Formulae. They represent the average correlation from all possible split half reliability estimates.

$$\text{K R 20} \quad r_{xx} = \frac{n}{n-1}\left(1 - \frac{\Sigma pq}{\sigma_o^2}\right)$$

$$\text{K R 21} \quad r_{xx} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{n\,\sigma_o^2}\right)$$

Where
- n = number of items in test
- p = proportion of students who answered item correctly
- q = proportion of students who answered item wrongly = (1 - p)
- pq = variance of a single item scored dichotomously (right/wrong).
- σ_o^2 = variance
- $\bar{x}$ = mean of the total test.

KR 21 assumes all items to be of equal difficulty (p is constant for all items). Given

this assumption, KR 21 is simply an algebric derivation of KR 20. If the assumption is incorrect, KR 21 will give slightly lower estimate of reliability.

KR 21 is a formula that classroom teachers can use and obtain estimates very easily. It requires very little computation. One needs to compute only the mean and variance of test scores.

Coefficient Alpha (α) Developed by Cronbach (1951) is a generalisation of KR 20 formula when the items are not scored dichotomously.

$$\alpha = \frac{n}{n-1} \quad 1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_0^2}$$

Where σ_i^2 = Variance of a single item

Hoyt's Analysis of Variance

This yields exactly the same results as KR - 20, stanley's approximation enables reliability estimates to be made both of objective type & choice type examination.

Scorer Reliability:

This is the marker reliability. In the case of objective tests, there is no need to find marker reliability. It must be 1. 0. But in the case of essay type, marker reliability will have to be ascertained. If a sample of papers has been scored independently by two different markers, the traditional' Pearson Product Moment Correlation Coefficient' (r) can be used to estimate the reliability of a single marker's scores. If one wishes to know the reliability of the sum (or average) of the reader's scores, he could use the Spearman Brown Formula. If more than two markers are used there are various intra class correlation formulae, one can use to obtain estimates of the scorer reliability of the summed scores. These require analysis of variance procedures.

Factors Influencing Reliability

There are many factors affecting reliability. Some of them are as follows:

1. Test Length:

When discussing the split-half method of estimating reliability, a specific case of the Spearman Brown Formula was illustrated. The more general expression of this formula is:

$$r_{\overline{xx}} = \frac{K_r}{1 + (K-1)r}$$

Where r_{xx} = predicted reliability of a test K times as long as original test.

r = reliability of original test

k = ratio number of items is new test to number of items in original one.

As an example, an illustrative example of 20 items Multiple-choice test had a reliability of 0.62. If we make it into a 60 items Multiple-choice test, this new test will have a reliability:

$$r_{xx} = \frac{3 \times 0.62}{1 + 2 \times 0.62} = \frac{1.86}{2.24} = 0.83$$

This assumes that the additional 40 items are equivalent to the earlier 20. We can use in this formula value of k = 1/2 or k = 1/3 to find reliabilities of tests of reduced lengths as well.

2. Speed

A test is considered a pure 'speed' test if everyone who reaches an item gets it right but no one has time to finish all the items. Thus, score differences depend upon the number of items attempted. The opposite of a speed test is a power test. A pure power test is one in which every one has time to try all items but because of the difficulty level, ordinarily no one obtains a perfect score. If a test is speeded, reliability should be computed by one of the methods that requires two administrations of the test.

3. Group Homogenity

The more heterogenous the group, the higher the reliability. $r_{xx} = 1 - \frac{V_e}{Vo}$. Here Vo increases with group heterogenity while Ve almost remains constant. Supposing a test is given to the entire first degree students first and then the same test administered on third year of first degree students it would be safe to conclude that because the third year first degree students are more homogenous, the reliability of the test for the third year students would be considerably lower than the reported reliability.

4. Difficutly of Items

The difficulty of the test and the individual items also affects the reliability of the test. Tests in which there is little variability among the scores give lower reliability estimates than tests in which the variability is large. Tests that are so easy that almost everyone gets all items correct or conversely, so hard that almost everyone gets all the items wrong will have little variability among the scores and will tend to have lower reliability.

5 Objectivity

Scorer reliability is high for objective type tests while the more subjectively a measure is scored the lower the reliability of the measure.

Reliability of Difference Scores

If two student's scores are 'different', we will be interested to know how appropriate is this observed difference compared to 'true' difference. Unfortunately, difference scores are considerably less reliable than single scores. The errors of measurement on each test contribute to error variance in the difference scores and the true variance that the two tests measure in common reduces the variability of the difference scores. Theoretically the reliability of the difference scores is the ratio of two variances. In this case, reliability is equal to the true variance of the difference scores divided by the observed variance of the difference scores.

If two tests have equal variances, the reliability of a difference score is the ratio of two variances. In this case, reliability is equal to the true variance of the difference scores divided by the observed variance of the difference scores.

If two tests have equal variances, the reliability of a difference score can be computed as:

$$r_{diff.} = \frac{\frac{r_{xx} + r_{yy}}{2} - r_{xy}}{(1 - r_{xy})}$$

Where $r_{diff.}$ = reliability of the difference scores

r_{xx} = reliability of one measure

r_{yy} = reliability of other measure

r_{xy} = correlation between the two measures.

S. E. of difference scores

The standard error of measurement of difference scores is

$$Se\ diff. = \sqrt{(SEM_x)^2 + (SEM_y)^2}$$

Where $(SEM_x)^2$ = Squared standard error of measurement for the x measure

$(SEM_y)^2$ = Squared standard error of measurement for the y measure

$Se\ diff. = Sx\sqrt{2 - r_{xx} - r_{yy}}$

Where $S_x = S_y$ = S.D. of either x or y measure.

The standard error of the difference is used when one wishes to decide whether an observed difference is of sufficient magnitude to be considered reliable or whether the observed difference could have occured by chance. Suppose a student has a score of 62 on verbal reasoning and score of 50 on mechanical reasoning subject of DAT, the estimated reliability of the two being 0.88, the standard deviation being 10, then

$$\text{Se.diff.} = 10/2 - 0.88 - 0.85 = 5.20$$

This Se difference of 5.20 means that we can be 68% confident that any difference between two subtests greater than 5.20 is a true difference and not due to chance alone.

Illustrative Example of a Test Analysis

Subject Matter of Test and Nature	:	Summative test on 'educational measurement' at the end of a semester work.
Number of students	:	76
Number of Items	:	20
Time	:	30 minutes (Almost a power test; everyone had the time to complete the test)
Maximum Marks	:	20
Test Constructor	:	V. Natarajan

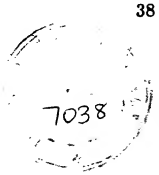
About the test

The test was a section of a total achievement test at the end of the semester. All the 20 items were of multiple choice type with four options, one of which was the correct answer. The students answered with the alphabet A. B. C. D. whichever was the answer to the item, put in a box in a separate response sheet provided. The question books together with the response sheets were collected.

Statistical analysis

Since there are 76 students, 1/6 of it will be 1/6 x 76 = 12.67 = 13 students' marks.

$$\text{S.d.} = \frac{(72 + 68 - 80) - (6 + 7 + 16 + 54 + 30)}{38}$$



$$= \frac{(220 - 113)}{38} = \frac{107}{38} = 2.81 \text{ against } 2.79 \text{ (accurate value)}$$

$$\text{Standard error of the mean} = \frac{\text{S.d.}}{\sqrt{N - 1}} = \frac{2.79}{\sqrt{75}} = 0.32$$

Mean of scores will lie between (12.9 - 0.32) and (12.9 + 0.32) for every 2 out of 3 cases approximately.

Reliability of the Test

The test of 20 items is split into two halves namely.

i) a test of odd numbered items and a test of even numbered items. Marks obtained by students on odd numbered item test and even numbered item test are all found out. Product moment correlation is worked out to give the split halves reliability. Spearman - Brown formula is used to find the whole test reliability from the split halves reliability.

ii) another way to have two halves of the same 20 item test is to take the first 10 item into a test & the last 10 as another. Students' marks on these first 10 and last 10 item tests are found and correlated. Spearman - Brown formula is used to find the whole test reliability from split halves reliability

iii) yet another way of making two tests out of one is to take any random 10 and constitute into a test while the rest will be made into another. Students' marks on these two tests are found and product moment correlation found out. Spearman-Brown formula is used to find the whole test reliability from this split halves.

To find out the value of r, the following formula is used.

$$r = \frac{\Sigma xy - \frac{\Sigma x \ \Sigma y}{N}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{N}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{N}}}$$

The value of r = .47

Applying spearman - Brown formula for whole test reliability from split halves = .6394

Rulon Formula

Rulon has developed a simple formula for reliability of total scores that follows closely the basic definition of reliability - that reliability is the proportion of true variance in a test. Rulon's equation however actually expresses the complementary statement that reliability is equal to unity minus the proportion of error variance.

$$r_{tt} = 1 - \frac{\sigma^2 d}{\sigma^2 t}$$

Where d = difference between two half scores for an examinee

σd = S.D. of those difference

σt = S.D. of total scores

$$r_{tt} = 1 - \frac{2.55}{7.79} = \frac{5.24}{7.79} = 0.6809$$

Flanagan Formula

Flanagan gives a formula parallel to Rulon's. It estimates the error variance in a sense as the sum of the variance of the two halves.

$$r_{tt} = \frac{2\,(1 - \sigma 1^2 + \sigma 2^2)}{\sigma t^2}$$

Where $\sigma 1^2$ & $\sigma 2^2$ = Variances of the two halves

σt^2 = Total variance

The value of r_{tt} for this test is = 0.672 Applying Spearman -Brown formula for the whole test reliability from the split halves = 0.6238.

Internal Consistency Reliability Estimates

i) Kuder - Richardson Formula 20:

In order to develop another approach to reliability that of internal consistency, it is sensible to consider the split half method as a philosophy. By using a split half method, the performance on different halves of an examination is compared.

$$KR\ 20 = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} (p_i q_i)}{\sigma o^2}\right)$$

Where n = number of items in the test

p_i = proportion of candidates answering the ith item correctly

q_i = $(1 - p_i)$

σo^2 = Variance of observed scores.

The results obtained by this formula are as under:

1. Using the entire population = 0.546
2. Using top and bottom 27% of population = 0.530

Tucker's modified K.R. Formula

To meet the demand for a simplified K.R. formula 20 and yet avoid the inaccuracies of K R 21 Tucker has developed a modified K R formula.

$$r_{tt} = \frac{n}{n-1}\left(\frac{\sigma_t^2 - n\bar{p}\bar{q} + n\sigma_p^2}{\sigma_t^2}\right)$$

Where σ_p = S D of the proportions of correct responses

$$\sigma_p^2 = \frac{\Sigma p^2}{n} - \bar{p}^2$$

Where $\bar{p}$ = mean of the proportion of correct responses for all items

The value of r_{tt} obtained by this formula is = 0.5448

Analysis of Variance Approach to Reliability

It is not surprising that the estimation of reliability can be made by a more conventional analysis of variance approach. Several investigators have proposed this kind of approach among whom are Jackson, Hoyt, and Alexander. Like the K.R. approach this one starts from the item level.

Hoyt basic formula for reliability is:

$$r_{tt} = \left(1 - \frac{V_r}{V_e}\right) = \frac{V_e - V_r}{V_e}$$

Where V_r = Variance for remainder sum of squares.

V_e = Variance for examinees.

The various sums of squares computed for this test of 20 items on 76 students is given below:

Source of variance	sum of squares	Degrees of freedom	Variance
Examinees	29.54	75	0.380
Items	69.22	19	-
Remainder	249.68	1425	0.170

The value of r_{tt} obtained by this formula is = 0.55 (same as KR 20)

Stanley's Procedure

We have seen that the reliability coefficient for an objective test requires extensive computation. J.C. Stanley(1964) has described a shortened method based on Kuder Richardson form 20 which is useful where item analysis is based on 27% top and bottom groups. Such a procedure is described by Melutosh and Morrison (1969) and the reliability is stanley's approximation which may be written as:

$$\text{Rel. Coeff. } (r) = \frac{K}{K-1} \quad 1 - \frac{2n \sum(N_H + N_L) - \sum(N_H + N_L)^2}{0.667 \times (N_H - N_L)^2}$$

Where K = number of test items

n = number of candidates in top or bottom 27% group

N_H = number of correct responses in top group

N_L = number of correct responses in lower group

This stanley's approximation formula yields a value for this 20 item objective type test a reliability coefficient of 0.546. A summary of the results given earlier is shown in the following table:

Summary of Results

S. No.	Type(or name of Reliability Coefficient	Value	Remarks
1.	Split half reliability odd-even	0.6394	0.518 $r_{1/21}$ Applying S-B formula r_{wt} = 0.684
2.	Mosier formula	0.5840	Short cut computing formula
3.	Rulon Formula	0.7060	Total score reliability
4.	Flanagan formula	0.6720	Total score reliability (simplest in family)
5.	Split half first half-second half	0.6238	
6.	Split half random 10-other 10	0.7125	
7.	KR-20 27% Upper-lower	0.5300	low value indicate that the conditions for KR formula are not satisfied, items are of widely differing facility; Test is short useful only for power tests.
	KR-20 whole group	0.5460	
8.	Coefficient Alpha (Cronbach)	0.6680	average of all possible split halves
9.	Stanley's approximation	0.5500	
10.	Tucker's modified K.R. formula	0.5480	
11.	Analysis of variance approach	0.5500	Same as K.R. 20

Discussion: It is felt that the lower bound value here is 0.55 and the highest value estimated is 0.68 leaving or two values exceeding this. Therefore, it is taken that an average value of 0.68 and 0.55 is 0.615 or 0.62. If we want to achieve a reliability of say 0.9, the length of the test can be increased by:

$$\frac{0.90\ (1-0.62)}{0.62\ x\ (1-0.90)} = \frac{9\ x\ 0.38}{0.62} = \frac{3.42}{0.62} = \text{(say 5 times)}$$

Reliability and its use:

What is an acceptable value of reliability for a test to take decisions about individuals or about groups? Usually recommended values are 0.85 and 0.65 respectively.

For general aptitude tests, stability estimate of reliability is very important For multiple aptitude tests, it is essential to have data on the reliabilities of the subtests and the difference scores. Equivalence and internal consistency estimates are also of value for interpreting, any Aptitude test. For achievement tests, equivalence reliability estimates, seem almost essential. Internal consistency estimates also should be provided.

A COMPARATIVE STUDY OF RELIABILITY OF CHOICE TYPE EXAMINATIONS

There are two related attributes that a good test or examination must possess. It must be reliable and valid. A valid test or examination is one that measures what it was intended to measure. A reliable test or examination is one that almost shows the same degree of consistency between two measures of the same thing. It is important to be clear about the relationship between these two attributes, which is not a reciprocal one. An examination can not be valid unless it is reliable;but it can be reliable without being valid. We must therefore consider these two characteristics separately. The reliability of the objective type tests can very easily be calculated rather than the choice type examinations. In measuring the reliability of objective tests this is usually done by assigning the odd items to one half and the even items to the other. Secondly it is necessary to allow for the fact that the correlation one obtains from this comparison is based on tests that are shorter than the original. Various methods are used to find out the reliability of choice type examinations. However,the best methods used for the same are given as follows:

1. The modified formula for Stanley's approximation to the Kuder Richardson Coefficient or reliability is in a form which may be applied generally to both objective and traditional essay type examination where FV = DI for all the items/questions are known K is the number of items on the objective test. In the case of a traditional essay type paper, K is the total number of questions that have been attempted i. e. all the questions on the paper.

$$r = \frac{k}{k-1}\left[1 - \frac{6\left[\Sigma (F/100 - \Sigma (F/100)^2\right]}{\left[\Sigma DI\right]^2}\right]$$

For a choice type examination paper with 8 Questions the value of

K = 8, K - 1 = 7

QN. No.	I	II	III	IV	V	VI	VII	VIII
F.V. (F/100)	0.625	0.540	0.585	0.265	0.120	0.440	0.230	0.48
D.I. (D)	0.48	0.75	0.52	0.50	0.24	0.50	0.49	0.46
$(F/100)^2$	0.3906	0.2916	0.3412	0.702	0.014	0.1936	Q 048	0.2304

$$\Sigma (F/100) = 3.275$$

$$\Sigma (F/100)^2 = 1.581$$

$$\Sigma (F/100) - \Sigma (F/100)^2 = 3.275 - 1.581 = 1.694$$

$$\Sigma DI = 3.94$$

$$(\Sigma DI)^2 = (3.94)^2$$

$$r = \frac{8}{7} \quad 1 - \frac{6 \times 1.694}{(3.94)^2} = \frac{8}{7} \quad \frac{(1 - 10.164)}{15.500}$$

Here we are applying it to choice type and therefore we can not depend upon the value

$$= \frac{8}{7} \quad \left(\frac{5.34}{15.50}\right) \quad = 0.394$$

2. Mean Question Inter Correlation Method

Here the approach is to consider the "variable N" question inter correlations, since it may be appreciated that if the examination is highly internally consistent then these correlations will be as high; conversely if the examination questions are measuring different traits these correlations will be very variable and it is possible even that some is negative. Thus it is reasonable to consider using these correlations in some way to estimate the examination mark reliability (internal consistency). Such a method would agree with the basic concepts of internal consistency. The first step in the method is to obtain the 'mean question inter correlation". This is achieved by using a transformation of correlation coefficients (to Fisher's Z) and then finding a mean value of a Z weighted by the numbers of candidates attempting various pair of questions. The transformation is necessary since correlation coefficients are non-linear in nature. After the averaging process the mean Z is converted back to a correlation coefficient the mean question intercorrelation.

This correlation may be regarded as internal consistency reliability of an examination of one question if all the questions in the examination paper were parallel, then this would be a parallel forms reliability estimate for one question. The problem is that we are not interested in an examination of only one question but one composed of requisite number of questions (K) whatever that may be. By using the Spearman - Brown formula it is possible to convert this unit question reliability into an estimate of the reliability of an examination K times longer. If we call $\bar{r}ij$ as the mean question inter correlation and r, the exam. reliability

$$\text{Then } r = \frac{k \quad \bar{r}_{ij}}{1 + (K - 1) \; r_{ij}}$$

Mean question inter correlation method

	Inter Correlation	Fisher's Z	Weight	Weighted Z
r_{12}	0.470	0.4382	88	38.560
r_{13}	0.470	0.4382	102	44.700
r_{14}	0.630	0.5581	108	60.270
r_{15}	0.5770	0.5205	4	2.082
r_{16}	0.270	0.2636	113	29.780
r_{17}	0.360	0.3452	42	14.500
r_{18}	0.310	0.3004	117	35.140

Mean Fisher's Z = 0.392

Correlation Coefficient $\bar{r}_{ij}$ = .414

$$r = \frac{K\ \bar{r}_{ij}}{1 + (K-1)\ \bar{r}_{ij}} = \frac{8 \times 0.414}{1 + 7 \times 0.414} = \frac{3.312}{3.898} = 0.85$$

Calculation of inter correlation between Q No. I marks and Q No. V marks.

Q No. I Marks x	Q No. V Marks y	x^2	y^2	xy
13	3	169	9	39
11	0	121	0	0
11	0	121	0	0
11	3	121	9	33

$$r_{15} = \frac{\Sigma xy - \frac{\Sigma x \ \Sigma y}{N}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{N}} \ \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{N}}}$$

$$= \frac{72 - \frac{46 \times 6}{4}}{\sqrt{532 - \frac{2116}{4}} \ \sqrt{18 - \frac{36}{4}}}$$

$$= \frac{3}{\sqrt{3 \times 3}} = \frac{1}{\sqrt{3}} = 0.577$$

reference to table for Fisher's Z, Z Value

for 0.577 = 0.5205

Similar inter correlations are calculated:

$r_{12} = 0.470$

$r_{13} = 0.470$

$r_{14} = 0.630$

$r_{15} = 0.5770$

$r_{16} = 0.270$

$r_{17} = 0.360$

$r_{18} = 0.310$

Cronbach's Coefficient Alpha

We can consider the most popular 5 questions out of 7 questions by the population. A combination of Q No. I (117), Q No. II (88), Q No. III (102), Q. No. IV (108), Q No. VI (113) and Q No. VIII (117) can be taken as an examination where all questions are compulsory. The number of candidates to be considered is 64 throughout. For those 64 students who have answered QI, II, III, IV, VI and VIII. it will be possible to find out Cronbach's Coefficient. This of course will be a lower bound value for reliability.

Q. No.	I	II	III	IV	VI	VIII
Variance of marks on this QN by 64 students	3.60	1.45	1.72	1.14	1.52	1.14

Cronbach's Coefficient

$$r = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_o^2}\right)$$

Where n = 6; $\sum_{i=1}^{6} \sigma_i^2 \quad (\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_6^2 + \sigma_8^2)$

$= 10.57$

σ_o^2 = Variance of scores (for these 64 candidates) = 29.93

$$r_\alpha = \frac{6}{5}\left(1 - \frac{10.57}{29.93}\right) = \frac{6}{5} \times \frac{19.36}{29.93} = 0.777$$

Formula N_4

This is proposed by Nuttal (1969). This of course is supposed to be the best method when there is no choice and all questions are compulsory.

$$r = \frac{1}{1-1}\left(1 - 1\left(\sum_{i}^{k} n_j S_j^2\right) \Big/ \left(S_x^2 \sum_{i}^{k} n_j\right)\right)$$

Where 1 = number of questions to be attempted = 6

k = number of questions set = 8

n_j = number of students who answer question j

S_j^2 = Variance of scores on question j

S_x^2 = Variance of total scores = 51.34

QN. No.	I	II	III	IV	V	VI	VII	VIII
No. answering the QN n_j	117	88	102	108	4	113	42	117
Variance on question j $(i_j)^2$ Scores	3.43	2.13	1.59	1.29	2.25	2.15	3.52	1.22
$n_j \; S_j^2$	390.78	188	162.18	139.32	9.0	242.95	147.84	142.74

$$\sum_{i=1}^{k} n_j \; S_j^2 = 1422.81 \qquad \sum_{i=1}^{k} n_j = 691$$

$$\therefore \quad r = \frac{6}{5} \left[1 - 6 \times 1422.8 / 51.34 \times 691 \right]$$

$$= 1.2 \times 0.76 = .912$$

Formula N_5 This is also proposed by Nuttal (1969)

$$r = \frac{1 \; (\bar{r}_{it})}{1 + (1 - 1) \; (\bar{r}_{it})^2}$$

1 = number of questions to be attempted = 6

$\bar{r}_{it}$ = mean correlation between question scores and total scores.

QN	I	II	III	IV	V	VI	VII	VIII
Correlation between question score and total score	0.65	0.60	0.82	0.62	0.97	0.72	0.68	0.54

r_{it} = Mean correlation between question scores and total scores

$$r = \frac{6 \times (0.694)^2}{1 + 5\,(0.694)^2} = \frac{2.8896}{3.4080} = 0.8478$$

KR - 21 Formula : Applying KR 21, r may be calculated by the following formula:

$$r = \frac{k}{k-1}\left[1 - \frac{\bar{x}\,(k - \bar{x})}{\sigma^2}\right]$$

$$= \frac{k}{k-1}\left[1 - \frac{(1 - \frac{\bar{x}}{k})}{\sigma^2}\right]$$

$$= \frac{6}{5}\left[1 - \frac{40.93 \times 59.07}{100 \times (7.15)^2}\right]$$

$$= 1.2\left[1 - \frac{40.93 \times 59.07}{100 \times 51.12}\right]$$

$$= 1.2\left[1 - \frac{2419}{5100}\right] = 1.2\left[\frac{2681}{5100}\right] = \frac{3217.2}{5100} = 0.63$$

$$r = 0.63$$

Formula P.

This is suggested by J.K. Backhouse (1972). This reduces to coefficient alpha when there is no choice of questions and all questions are compulsory. The method of derivation is analogous to K R 20 given by Gullikson (1950)

$$r_{xy} = (\lambda + 1)\left(1 - \frac{\sum_{j=1}^{k} n_j S_j^2}{n S_x^2}\right) - \frac{\sum_{j,t=1}^{k} n_{j,t}\, m_{j,t}\, m_{t,j} - n M_x^2}{n S_x^2}$$

Where $\lambda = \left(\sum_{j=1}^{k} n_j\right) \Big/ \left(\sum_{j,t=1}^{k} n_{j,t}\right)$

k	=	number of questions in the test
n	=	number of people taking the test
X_{ij}	=	Score of ith person on jth question
X_i	=	Score of ith person on the whole test
m_j	=	mean of scores on the jth question
n	=	number of people attempting jth question
S_x^2	=	variance of test scores
$n_{j,t}$	=	number of people attempting both questions j and t.
$r_{j,t}$	=	correlation coefficient between scores on questions j and t.
$m_{j,t}$	=	mean scores on question j of these who also answered question t.
S_{jjt}^2	=	variance of scores on question j of those who also answered question t

For a detailed study of the reliability of choice-type examination the author refers the readers to go through the 'Monograph on Test and Item Analysis for Universities published by the Association of Indian Universities, Rouse Avenue, New Delhi-110002.